Research Article

Psychological Science
2018, Vol. 29(11) 1807–1823

DOI: 10.1177/0956797618788882
www.psychologicalscience.org/PS

Inferring Whether Officials Are Corruptible From Looking at Their Faces

Chujun Lin, Ralph Adolphs, and R. Michael Alvarez
Division of Humanities and Social Sciences, California Institute of Technology

Abstract

While inferences of traits from unfamiliar faces prominently reveal stereotypes, some facial inferences also correlate with real-world outcomes. We investigated whether facial inferences are associated with an important real-world outcome closely linked to the face bearer's behavior: political corruption. In four preregistered studies (N = 325), participants made trait judgments of unfamiliar government officials on the basis of their photos. Relative to peers with clean records, federal and state officials convicted of political corruption (Study 1) and local officials who violated campaign finance laws (Study 2) were perceived as more corruptible, dishonest, selfish, and aggressive but similarly competent, ambitious, and masculine (Study 3). Mediation analyses and experiments in which the photos were digitally manipulated showed that participants' judgments of how corruptible an official looked were causally influenced by the face width of the stimuli (Study 4). The findings shed new light on the complex causal mechanisms linking facial appearances with social behavior.

Keywords

face perception, corruption, social attribution, stereotyping, political psychology, open data, open materials, preregistered

Received 9/25/17; Revision accepted 6/3/18

Faces are rich in information: They provide clues about gender, race, age, and trait attributes, which are inferred spontaneously and ubiquitously (Engell, Haxby, & Todorov, 2007; Todorov, 2017). Moreover, such inferences often guide our social behavior—for instance, we decide whom to trust on the basis of how trustworthy a face looks (Rezlescu, Duchaine, Olivola, & Chater, 2012; Van't Wout & Sanfey, 2008). Many trait judgments made by participants across generations and cultures show consensus (Cogsdill, Todorov, Spelke, & Banaji, 2014; Lin, Adolphs, & Alvarez, 2017; Rule et al., 2010). But are trait judgments from faces accurate?

Previous research has shown that trait judgments from faces can be associated with important real-world social outcomes, such as dating and mating (Olivola et al., 2014; Valentine, Li, Penke, & Perrett, 2014), earnings and fundraising (Genevsky & Knutson, 2015; Hamermesh, 2011; Ravina, 2012), science communication (Gheorghiu, Callan, & Skylark, 2017), sentencing decisions (Berry & Zebrowitz-McArthur, 1988; Blair, Judd, & Chapleau, 2004; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), and leader selection (Todorov, Mandisodza, Goren, & Hall, 2005; for reviews, see Antonakis & Eubanks, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Yet this prior research on the association between trait judgments from faces and real-world outcomes leaves open two important questions. First, most associations have focused on prosocial outcomes (e.g., correlations between competence judgments and election success; Todorov et al., 2005). Second, most associations are plausibly driven not by the behavior of the targets whose face is being judged but by the interests of the perceivers who are making the judgments (e.g., correlations between interesting-looking scientists and the perceiver's interest in their work). Here, we investigated an antisocial judgment that

Corresponding Author:
Chujun Lin, California Institute of Technology, Division of Humanities and Social Sciences, HSS 228-77, 1200 E. California Blvd., Pasadena, CA 91125
E-mail: clin7@caltech.edu

may offer a clearer insight into associations with the judged person's own behavior: political corruption.

Political corruption has been a major cause of regime change and an important subject of much study in political science and economics (Rose-Ackerman, 2013). The possibility that corruptibility inferences from faces might be associated with real-world measures of corruption is raised by three areas of previous research. First, theories of self-fulfilling prophecy argue that the impressions and expectations a face creates (e.g., how corruptible an official looks) influence how other people interact with the face bearer (e.g., how likely others would be to bribe the official) and that those recurrent interactions in turn shape the face bearer's behavior so as to confirm other people's impressions and expectations (Haselhuhn, Wong, & Ormiston, 2013; Jussim, 1986; Slepian & Ames, 2016). Second, analyses of sentencing decisions show that evaluations of guilt and recommendations of punishment are influenced by the defendant's facial appearance (Berry & Zebrowitz-McArthur, 1988; Blair et al., 2004; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991). These findings suggest that officials who look more corruptible might be more likely to be accused, prosecuted, and convicted. Third, some studies have argued that the face contains a kernel of truth about a person's nature—such as personality and criminal inclinations (Penton-Voak, Pound, Little, & Perrett, 2006; Valla, Ceci, & Williams, 2011)—even though the diagnostic validity and the causal mechanisms remain obscure.

Given past research, we hypothesized that elected officials' corruption records would be associated with traits, such as corruptibility, inferred from their facial appearances. We examined this association in three preregistered studies, where participants made trait inferences on the basis of the photos of unfamiliar government officials. To account for the possibility that this association might depend on the severity of corruption and the level of office, we inspected both serious violations (i.e., cases considered political corruption) and minor violations (i.e., cases meriting a fine) and included officials at different levels of government (federal, state, and local). In a fourth preregistered study, we explored which facial features might be causally mediating the impression of how corruptible an official was, using mediation analyses as well as experimental manipulations of the face stimuli. In this fourth study, we focused on metrics of facial structures—in particular, facial width (relative to facial height) because it has been reported that men with wider faces are judged as less trustworthy (Haselhuhn et al., 2013; Stirrat & Perrett, 2010), more threatening (Geniole, Denson, Dixson, Carré, & McCormick, 2015), and less than fully human (Deska, Lloyd, & Hugenberg, 2018), although it remains unknown whether facial width-to-height ratio associates with actual behavior.

We have reported all measures, all conditions, all data exclusions, and how sample sizes were determined in this article and on the Open Science Framework (https://osf.io/k4mds/). All materials, data, and analysis codes for the present research can be accessed at this link.

Study 1

Our first study focused on federal and state officials and compared those who had clean records with those who were convicted of political corruption.

Method

Participants. This study was preregistered before data collection began (https://osf.io/mge8r/). A sample size of 100 participants was predetermined on the basis of two pilot studies—one carried out in the lab in May 2016 and the other via Amazon's Mechanical Turk (MTurk) in October 2016. The lab study included 32 participants recruited from the general public of Southern California, and the MTurk study had 18 participants. For the hypothesis that elected officials' corruption records would be associated with face-based inferences of corruptibility, the laboratory pilot study yielded an estimated effect size of 1.06, and the MTurk pilot study yielded an estimated effect size of 1.05, justifying a minimum sample size of 16 participants. Given these results and to ensure sufficient power even with dropout, we recruited 100 MTurk participants in November 2016. We selected participants who were native English speakers, located in the United States, and 18 years old or older. In addition, they had to have normal or corrected-to-normal vision, an educational attainment of high school or above, a good MTurk participation history (a human-intelligence-task, or HIT, approval rate ≥ 95% and ≥ 1,000 HITs approved), and no prior participation in our pilot studies.

Eighteen individuals were excluded in total, 2 for not being native English speakers, 6 for pressing the same response key for all trials in a block, and 10 for failing to input valid responses for more than 10% of the trials in a block (responses were considered not valid if missing or entered within 100 ms—the minimum time needed for visual exploration of the face; Olivola & Todorov, 2010). After exclusion, there were 82 participants in our final sample (42 female; age: $M = 39$ years, $SD = 12$; 84% White, 10% Black, 5% Asian).

Stimuli. Stimuli were photos of 72 real elected officials. All were Caucasian males who have held federal or state legislative offices in the United States. Photos were official headshots obtained from government websites and personal campaign websites (63%), news articles (23%), and Wikipedia (14%). All photos were converted to gray-scale

images on a plain gray background and cropped to a uniform size. All faces were frontal, smiling, in clear focus, and centered in the middle of the image.

Among the 72 officials, half were convicted of political corruption (*corrupt officials*), and the other half had clean records (*noncorrupt officials*). The corrupt officials were from two Wikipedia data sets (list of American state and local politicians convicted of crimes: https://en.wikipedia.org/wiki/List_of_American_state_and_local_politicians_convicted_of_crimes; list of American federal politicians convicted of crimes: https://en.wikipedia.org/wiki/List_of_American_federal_politicians_convicted_of_crimes). To reduce sources of variability, we included only officials who were Caucasian, were male, held federal or state legislative offices, and were convicted between 2001 and 2016 of political corruption conducted while in office (bribery, money laundering, embezzlement, mail fraud, wire fraud, tax fraud, conflict of interest, misusing funds, misusing office, or falsifying records). In addition, age information for these officials had to be publicly available, as did frontal photographs of acceptable clarity in which the official was smiling. All photographs had been taken while officials were in office. Most photos of the corrupt officials had a known creation date, and we confirmed that the photos were taken before their conviction (72%); for the rest of the photos (28%), the creation date was unknown (analyses were also performed when excluding data for these stimuli; the pattern of results did not change). The noncorrupt officials were randomly matched from the list of incumbents who had clean records, were holding the same office in the same state, and were of the same gender, the same race, and similar age (±12 years) as the corrupt officials during the period of their misconduct. For instance, if the stimuli contained a Caucasian male corrupt official who was a member of the Arizona House of Representatives during his misconduct at the age of 55, then a noncorrupt official would be randomly selected from our available stimulus set from the list of Arizona House of Representatives incumbents who had a clean record and who was a Caucasian male between the ages of 43 and 67.

Procedures. Participants were not informed of the purpose of the study or the sampling of the stimuli. In particular, they were not given any information about the percentage of politicians in our stimulus set who might be corrupt in real life. They were told only that they would view a series of politician photos and that they should judge how corruptible, dishonest, selfish, trustworthy, and generous these politicians looked to them (experiment instructions are available at https://osf.io/k4mds/). Participants completed five blocks of experiments, with each block corresponding to judging one trait for all faces. The ordering of the faces within each block as well as the ordering of the blocks were randomized.

Each block started with an instruction screen that specified the trait to be judged (e.g., corruptibility). Participants were instructed to make their decisions as quickly and precisely as possible. Six practice trials familiarized participants with the task. Participants viewed photos of officials one at a time in randomized order and made judgments. Each trial began with a fixation cross, followed by the photo (1 s) with a 5-point Likert scale below it. Scales were anchored with bipolar adjectives (Fig. 1). Participants could make a decision as soon as the photo appeared and within 4 s after the photo disappeared. The orientation of the scale was randomized across blocks, and scores were reverse-coded as needed.

After completing all five blocks of ratings, participants were asked whether they had recognized any of the officials and filled out a short survey questionnaire

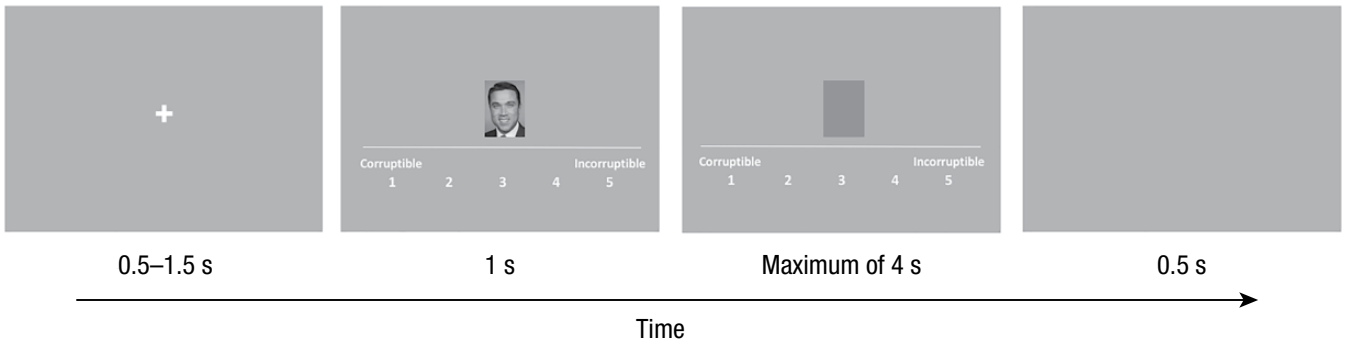

Fig. 1. An example trial in the corruptibility-judgment block. Each trial began with a fixation cross. Then a photo of an official appeared for 1 s. The orientation of the scale was randomly flipped for each block and each participant. Participants made a decision by pressing one of the number keys from "1" to "5" on their keyboard. As soon as a valid key was pressed (or 4 s after the photo disappeared if no valid key was pressed), the trial ended, and there was a blank interstimulus interval.

on demographic characteristics, political attitudes, and personality.

Results

Reliability of face-based trait inferences. Following our preregistered plan, we excluded from further analysis any responses faster than 100 ms and data for officials who were recognized. Among the 82 participants, 7 recognized one official (in total, four officials were recognized). The percentages of participants who used the full scale to rate the faces ranged from 59% to 68% across the five traits, and over 90% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S1 in the Supplemental Material available online).

First, we checked whether participants gave consistent judgments for a face across different traits. We expected consistent ratings for a face on traits with the same valence to be positively correlated and ratings on traits with opposite valences to be negatively correlated. Although this was not planned in our preregistration, we computed repeated measures correlations (the R function *rmcorr*) to determine the common within-individuals correlations for ratings between each pair of traits, to handle the issue of nonindependence in repeated measures. Results (see Table S1 in the Supplemental Material for coefficients and 95% confidence intervals, or CIs) showed that at an individual level, judgments of a face for traits with the same valence were positively correlated (repeated measures rs ranging from .24 to .31, ps < .001), and judgments of a face for traits with opposite valences were negatively correlated (repeated measures rs ranging from –.30 to –.21, ps < .001). Following our preregistered plan, we also analyzed the consistency of these ratings at an aggregate level. Ratings for each face were first averaged over participants, and (tie-corrected) Spearman correlation coefficients were calculated for each pair of traits with those averaged ratings. Aggregate-level judgments for a face were highly consistent across traits because they averaged out the measurement noise inherent in the individual-level correlations ($|r| \geq .75$; see Table S2 in the Supplemental Material).

Next, intraclass correlation coefficients (ICCs) were computed for each trait separately to test whether inferences of a trait showed consensus across participants—ICCs were computed according to type ICC(2, k) on the basis of complete cases. A high ICC indicates that the total variance in ratings is mainly explained by rating variance across images instead of across participants. In line with prior literature (see the introduction), our results showed high consensus among participants for inferences of corruptibility, ICC = .81, $F(48, 3888) = 6.4$, 95% CI = [.73, .88]; dishonesty, ICC = .82, $F(45, 3645) = 6.7$, 95% CI = [.74, .89]; selfishness, ICC = .86, $F(42, 3402) = 8.1$, 95% CI = [.80, .91]; trustworthiness, ICC = .82, $F(43, 3483) = 6.7$, 95% CI = [.74, .89]; and generosity, ICC = .82, $F(43, 3483) = 6.6$, 95% CI = [.74, .89].

Association between corruption records and face-based trait inferences: preregistered analyses. Our primary interest in the current study was the extent to which trait inferences from a face were associated with actual corruption records. First, we followed the analysis methods planned in our preregistration and tested for these associations on the basis of inference judgments aggregated across participants and individually within subjects. For inferences of a negative trait, we deemed an official to be categorized accurately if he was convicted of corruption and received a high rating (> 3) or, conversely, if he had a clean record and received a low rating (≤ 3); for inferences of a positive trait, we deemed an official to be categorized accurately if he was convicted of corruption and received a low rating (< 3) or, conversely, if he had a clean record and received a high rating (≥ 3).

One-sample, one-tailed proportion tests against chance (50%) were performed on the aggregated-level accuracies across officials. One-sample one-tailed t tests against chance (50%) were performed on the individual-level accuracies across participants (we also calculated individual-level accuracies by categorizing midpoint 3 in the opposite way; see Table S3 in the Supplemental Material). Results (summarized in Table 1) showed that both aggregate-level and individual-level inferences of traits were associated with actual corruption records of the facial identities at a level better than chance (see Fig. S2 in the Supplemental Material for full distributions of individual-level accuracies).

Association between corruption records and face-based trait inferences: extensions to preregistered analyses. Beyond our planned preregistered analyses, we conducted three additional robustness checks on the association between trait inferences from faces and corruption records. First, we confirmed that the above-chance accuracy we observed was not driven just by a small subset of faces: For each trait, we ranked the officials by the number of participants who categorized them accurately; we then calculated the average individual-level accuracy for subsets of stimuli in which the officials were progressively excluded one by one from the official who was accurately categorized by most participants to the official who was accurately categorized by fewest participants. For all five traits, average individual-level accuracies decreased smoothly as the highest ranked officials were removed and stayed above chance even after the 12th

Table 1. Results for Correctly Categorized Officials Based on Aggregate-Level Trait Inferences and Individual-Level Trait Inferences From Study 1

	Aggregate-level accuracy				Average individual-level accuracy[a]				
Trait	Percentage of correctly categorized officials ($N = 72$)	Lower bound of 95% CI	$\chi^2(1)$	p	Mean accuracy ($N = 82$)	SD	Lower bound of 95% CI	$t(81)$	Cohen's d
Corruptibility	69.44%	59.22%	10.13	< .001	55.73%	6.95%	54.46%	7.47	0.82
Dishonesty	70.83%	60.67%	11.68	< .001	54.82%	6.41%	53.64%	6.81	0.75
Selfishness	66.67%	56.36%	7.35	.003	55.10%	6.76%	53.86%	6.83	0.75
Trustworthiness	68.06%	57.79%	8.68	.002	55.03%	6.41%	53.85%	7.10	0.78
Generosity	63.89%	53.53%	5.01	.013	54.97%	5.99%	53.87%	7.51	0.83

Note: CI = confidence interval.
[a]All ps for this variable are less than .001.

highest ranked official was excluded from the stimulus set (see Table S4 in the Supplemental Material).

Second, although participants were not informed of the purpose of the study or the percentage of corrupt politicians in our stimulus set (they were told only that these people were politicians), their beliefs (implicit or explicit) about the base rates of corrupt politicians in the real world or the percentage of corrupt politicians in our experiment might bias the ratings they gave. We corrected for such possibly idiosyncratic biases among our participants by calculating individual-level accuracies using an alternative method. Ratings for each participant were centered on that participant's mean across all of his or her ratings on a trait (see Fig. S3 in the Supplemental Material for the full distributions of mean ratings).

For this analysis, inferences of a negative trait were deemed accurate if the official had been convicted of corruption and received a rating from a participant that was higher than the participant's mean rating or, conversely, if the official had a clean record and received a rating from a participant that was lower than the participant's mean rating; inferences of a positive trait were deemed accurate if the official was convicted of corruption and received a rating from a participant that was lower than the participant's mean rating or, conversely, if the official had a clean record and received a rating from a participant that was higher than the participant's mean rating. One-sample, one-tailed t tests against chance (50%) were performed on individual-mean-centered accuracies across participants. Corroborating the results reported previously, individual-level trait inferences correlated with officials' corruption records at a level better than chance, and the effect sizes were large—corruptibility inferences: $M = 55.57\%$, $SD = 7.75\%$, lower bound of 95% CI = 54.14%, $t(81) = 6.50$, $p < .001$, $d = 0.72$; dishonesty inferences: $M = 55.12\%$, $SD = 6.43\%$, lower bound of 95% CI = 53.94%, $t(81) = 7.22$, $p < .001$, $d = 0.80$; selfishness inferences: $M = 54.95\%$, $SD = 7.87\%$, lower bound of 95% CI = 53.50%, $t(81) = 5.69$, $p < .001$, $d = 0.63$; trustworthiness inferences: $M = 55.59\%$, $SD = 6.53\%$, lower bound of 95% CI = 54.39%, $t(81) = 7.75$, $p < .001$, $d = 0.86$; and generosity inferences: $M = 55.31\%$, $SD = 6.95\%$, lower bound of 95% CI = 54.03%, $t(81) = 6.92$, $p < .001$, $d = 0.76$.

Third, to address the concern that dichotomizing ratings into accurate and inaccurate might lead to loss of measurement sensitivity and to handle the nonindependence in ratings due to repeated measures designs, we performed general linear mixed-model (GLMM) analyses for inferences of each trait, respectively. Officials' corruption records (1 = conviction, 0 = clean) were regressed on individual-level ratings in logistic models, and participants were treated as random factors ($N = 5{,}757$; N was determined by the number of participants multiplied by the number of faces, excluding omitted observations; observations from a participant for a face would be omitted if ratings were not available for all five traits). In addition, photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as control variables in all models. All continuous variables were standardized.

We observed significant effects of trait ratings: Officials who were rated as looking more corruptible, $b = 0.23$, $SE = 0.03$, 95% CI = [0.17, 0.29], $z = 7.66$, $p < .001$; dishonest, $b = 0.17$, $SE = 0.03$, 95% CI = [0.11, 0.23], $z = 5.75$, $p < .001$; and selfish, $b = 0.20$, $SE = 0.03$, 95% CI = [0.14, 0.26], $z = 6.77$, $p < .001$, were more likely to have been convicted of corruption, whereas officials who were rated as looking more trustworthy, $b = -0.19$, $SE = 0.03$, 95% CI = [−0.25, −0.13], $z = -6.41$, $p < .001$, and generous, $b = -0.20$, $SE = 0.03$, 95% CI = [−0.26, −0.14], $z = -6.59$, $p < .001$, were less likely to have been convicted of corruption (for complete lists of coefficients, see Table S5 in the Supplemental Material).

Association between corruption records and face-based trait inferences: further exploration of potential mechanisms. Finally, we performed two additional analyses that were also beyond our preregistration. We performed GLMM analyses on two subsets of data to test two photo-selection–related mechanisms underlying the face–corruption-record association we found. To test the hypothesis that potential negative biases in the convicted officials' photos that were from sources beyond the control of the officials might be driving the association, we conducted GLMM analyses on a subset of data that included only officials whose photos were self-selected—that is, those from government websites and personal campaign websites (n = 45; 20 were convicted of corruption; in this subset, only 1 official had a beard, and only 2 officials were bald, and therefore these two predictors were removed from the model).

The associations between trait inferences and records remained significant—corruptibility inferences: b = 0.24, SE = 0.04, 95% CI = [0.17, 0.32], z = 6.81, p < .001; dishonesty inferences: b = 0.19, SE = 0.04, 95% CI = [0.12, 0.26], z = 5.21, p < .001; selfishness inferences: b = 0.18, SE = 0.04, 95% CI = [0.11, 0.25], z = 5.07, p < .001; trustworthiness inferences: b = −0.20, SE = 0.04, 95% CI = [–0.27, –0.13], z = −5.63, p < .001; and generosity inferences: b = −0.17, SE = 0.04, 95% CI = [–0.24, –0.10], z = −4.66, p < .001.

To test the hypothesis that potential negative biases in the convicted officials' photos that were taken after conviction might be driving the face–corruption-record association, we conducted GLMM analyses on a subset of data that included only officials whose photo dates were known (and were prior to the date of conviction, for convicted officials; n = 62; 26 were convicted of corruption). The associations between trait inferences and records became weaker but remained significant—corruptibility inferences: b = 0.17, SE = 0.03, 95% CI = [0.10, 0.23], z = 4.93, p < .001; dishonesty inferences: b = 0.11, SE = 0.03, 95% CI = [0.04, 0.18], z = 3.29, p = .001; selfishness inferences: b = 0.16, SE = 0.03, 95% CI = [0.09, 0.22], z = 4.64, p < .001; trustworthiness inferences: b = −0.14, SE = 0.03, 95% CI = [–0.20, –0.07], z = −4.06, p < .001; and generosity inferences: b = −0.19, SE = 0.03, 95% CI = [–0.26, –0.13], z = −5.74, p < .001. This indicates that while potential biases in photo selection can explain some of the relationship between trait ratings and officials' records, they cannot entirely account for our main findings.

Two additional analyses were preregistered but are not presented in this article; the codes to conduct those analyses can be found at https://osf.io/k4mds/. In our preregistration, we proposed an alternative approach to analyze individual-level ratings (logistic regression with adjusting standard errors for clustering). These analyses are not presented here because the GLMM analyses reported previously are more appropriate for handling repeated measures. We had also planned analyses of correlations between individual-level accuracies and response times, but these were intended to answer a question that is beyond the scope of the current article.

Study 2

Study 1 showed that compared with peers with clean records, federal and state officials who were convicted of political corruption were perceived as more corruptible, dishonest, and selfish and less trustworthy and generous. To assess the generalizability of these findings, we next tested whether they would also hold for officials from lower levels of governments and for the comparison between officials with clean records and officials who violated campaign finance laws.

Method

Participants. This study was preregistered before data collection began (https://osf.io/tgzpz/). A pilot study with 24 MTurk workers conducted in February 2017 yielded an estimated effect size of 1.39, justifying a minimum sample size of 10 participants. To ensure sufficient power and to have a sample size comparable with that of Study 1, we predetermined the sample size to be 100 participants. The same inclusion and exclusion criteria as in Study 1 were applied (including exclusion of participants from Study 1). We excluded 22 individuals, 3 for not being native English speakers, 2 for pressing the same response key for all trials in a block, and 17 for failing to input valid responses for more than 10% of the trials in a block. After these exclusions, there were 78 MTurk workers who participated in this study in February and March 2017 (33 female; age: M = 38 years, SD = 11; 83% White, 9% Black, 6% Asian).

Stimuli. Stimuli were photos of 80 real elected officials. All officials were Caucasian males who held offices in California state and local governments. Photos were official headshots obtained from government websites and personal campaign websites (86%), news articles, and Wikipedia (14%). All photos were converted to gray scale on a plain gray background and were cropped to a uniform size. All faces were frontal, smiling, in clear focus, and centered in the middle of the image.

Among the 80 officials, half violated the California Political Reform Act (*officials with violations*), and the other half had clean records (*officials without violations*). The officials with violations were from the database of the California Fair Political Practices Commission's

"Enforcement Cases" (http://www.fppc.ca.gov/about-fppc/hearings-meetings-workshops/current-agenda/past-agendas.html). To reduce sources of variability, we included only officials who were Caucasian, were male, and had committed a violation related to election campaigns (laundered campaign contributions, accepted over-the-limit gifts and contributions, improperly used campaign funds, had conflicts of interest, inadequately or inaccurately reported on campaign statements, did not file campaign statements or filed them late, or were involved in illegal campaign coordination). In addition, we included only successful candidates of the election related to the violation, whose cases merited pursuit of a fine over $215, whose cases were closed between January 2015 and January 2017, whose age information was publicly available, and who had publicly available frontal photographs of acceptable clarity that featured them smiling. All photographs had been taken while in office. Most photos of the officials with violations had a known creation date, and we confirmed that the photos were taken before the cases were closed (88%); for the rest of the photos (12%), the creation date was unknown (analyses were also performed when excluding data for these stimuli). The officials without violations were randomly generated from our available stimulus set from the list of incumbents who had clean records and were holding the same office in the state of California and were the same gender, the same race, and of similar age as the officials with violations.

Procedure. Participants followed the same experimental procedure as in Study 1 but viewed a new set of stimuli, as described previously.

Results

Reliability of face-based trait inferences. Following our preregistered plan, we excluded responses faster than 100 ms and responses for officials who were recognized. Among the 78 participants, only 1 recognized one official. As in Study 1, ratings across faces given by each participant had sufficient variance: The majority of participants used the full scale to rate the faces (the percentages of participants ranged from 58% to 63% across the five traits), and more than 97% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S4 in the Supplemental Material).

To test how consistently a participant judged a face across different traits, we computed repeated measures correlations (using the R function *rmcorr*) following the method in Study 1. A participant's ratings of a face on traits with the same valence were positively correlated (repeated measures rs ranging from .26 to .35, ps < .001), and ratings on traits with opposite valences were negatively correlated (repeated measures rs ranging from –.38 to –.26, ps < .001; see Table S6 in the Supplemental Material for coefficients and 95% CIs). As planned in our preregistration, we also computed (tie-corrected) Spearman correlation coefficients for each pair of traits using ratings averaged over participants for each face. Aggregate-level judgments of a face were once again highly consistent across traits ($|r| \geq .77$; see Table S7 in the Supplemental Material).

In line with Study 1 and prior literature, we observed high consensus among participants for face-based judgments of corruptibility, ICC = .81, $F(64, 4928) = 6.3$, 95% CI = [.75, .87]; dishonesty, ICC = .82, $F(61, 4697) = 6.7$, 95% CI = [.75, .87]); selfishness, ICC = .82, $F(45, 3465) = 6.2$, 95% CI = [.74, .89]; trustworthiness, ICC = .86, $F(58, 4466) = 8.6$, 95% CI = [.81, .91]; and generosity, ICC = .87, $F(56, 4312) = 8.9$, 95% CI = [.82, .91]. ICCs were computed according to type ICC(2, k) on the basis of complete cases.

Association between records of violations and face-based trait inferences: preregistered analyses. Following the methods in Study 1, we calculated the proportions of correctly categorized officials for each trait on the basis of aggregate-level inferences and individual-level inferences as planned in our preregistration. Table 2 summarizes one-sample one-tailed proportion-test statistics of aggregate-level accuracies and one-sample one-tailed t-test statistics of individual-level accuracies (see Fig. S5 in the Supplemental Material for full distributions of individual-level accuracies; see Table S8 in the Supplemental Material for average individual-level accuracies calculated with categorizing midpoint 3 in an opposite way). The findings replicated those from Study 1.

Association between corruption records and face-based trait inferences: extensions to preregistered analyses. As in Study 1, we conducted three analyses in addition to those we had preregistered to check the robustness of the association between trait inferences from officials' faces and the records of violations of the facial identities. First, we verified that the above-chance accuracy observed earlier was not driven just by a small subset of faces. Following the same approach as Study 1, we recalculated individual-level accuracies for subsets of stimuli in which the stimulus was excluded one by one from the official who was accurately categorized by most participants to the official who was accurately categorized by the fewest participants. Average individual-level accuracies for each trait decreased smoothly as the highest ranked officials were progressively excluded and stayed above chance even after the 14th highest ranked official was excluded from the stimulus set (see Table S9 in the Supplemental Material).

Table 2. Results for Correctly Categorized Officials Based on Aggregate-Level Trait Inferences and Individual-Level Trait Inferences From Study 2

	Aggregate-level accuracy				Average individual-level accuracy[a]				
Trait	Percentage of correctly categorized officials ($N = 80$)	Lower bound of 95% CI	$\chi^2(1)$	p	Mean accuracy ($N = 78$)	SD	Lower bound of 95% CI	$t(77)$	Cohen's d
Corruptibility	67.50%	57.79%	9.11	.001	54.72%	6.59%	53.48%	6.32	0.72
Dishonesty	70.00%	60.38%	12.01	< .001	56.15%	6.51%	54.92%	8.34	0.94
Selfishness	65.00%	55.23%	6.61	.005	55.78%	7.21%	54.42%	7.08	0.80
Trustworthiness	70.00%	60.38%	12.01	< .001	56.00%	6.31%	54.74%	7.98	0.90
Generosity	67.50%	57.79%	9.11	.001	55.80%	5.51%	54.76%	9.29	1.05

Note: CI = confidence interval.
[a]All ps for this variable are less than .001.

Second, participants' beliefs (implicit or explicit) about the base rates of corrupt politicians in the real world or the percentage of corrupt politicians in our study might have influenced their trait judgments from politicians' faces. Consequently, we computed the individual-level accuracies using an alternative method that took into account the heterogeneous beliefs of base rates across participants. As in Study 1, a mean rating was computed for each participant by averaging the ratings he or she gave across all faces for a trait (see Fig. S6 in the Supplemental Material for the full distributions of mean ratings). This mean rating was used as a cutoff for dichotomizing whether the participant's rating correctly categorized an official. These individual-mean-centered accuracies across participants were then tested against chance (50%). We observed significantly above-chance accuracies and large effect sizes for corruptibility inferences, $M = 55.06\%$, $SD = 6.98\%$, lower 95% CI = 53.74%, $t(77) = 6.40$, $p < .001$, $d = 0.72$; dishonesty inferences, $M = 56.06\%$, $SD = 7.32\%$, lower 95% CI = 54.68%, $t(77) = 7.31$, $p < .001$, $d = 0.83$; selfishness inferences, $M = 55.74\%$, $SD = 7.98\%$, lower 95% CI = 54.24%, $t(77) = 6.36$, $p < .001$, $d = 0.72$; trustworthiness inferences, $M = 56.00\%$, $SD = 7.05\%$, lower 95% CI = 54.67%, $t(77) = 7.52$, $p < .001$, $d = 0.85$; and generosity inferences, $M = 55.61\%$, $SD = 6.62\%$, lower 95% CI = 54.36%, $t(77) = 7.48$, $p < .001$, $d = 0.85$.

Third, data were further analyzed in GLMM analyses to handle the nonindependence in ratings due to the repeated measures design and avoid any data dichotomization. Officials' records of violations (1 = violation, 0 = clean) were regressed on individual-level ratings in logistic models, and participants were treated as random factors ($N = 6{,}115$; N was determined by the number of participants multiplied by the number of faces excluding omitted observations; observations from a participant for a face would be omitted if ratings were not available for all five traits). In addition, photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as control variables in all models. All continuous variables were standardized. Results revealed significant effects of trait ratings: Officials who were rated as looking more corruptible, $b = 0.24$, $SE = 0.03$, 95% CI = [0.18, 0.29], $z = 8.19$, $p < .001$; dishonest, $b = 0.28$, $SE = 0.03$, 95% CI = [0.23, 0.34], $z = 9.77$, $p < .001$; and selfish, $b = 0.27$, $SE = 0.03$, 95% CI = [0.21, 0.32], $z = 9.31$, $p < .001$, were more likely to have violated campaign finance laws, whereas officials who were rated as looking more trustworthy, $b = -0.26$, $SE = 0.03$, 95% CI = [–0.32, –0.20], $z = -9.05$, $p < .001$, and generous, $b = -0.27$, $SE = 0.03$, 95% CI = [–0.33, –0.22], $z = -9.53$, $p < .001$, were less likely to have violated campaign finance laws (for complete lists of coefficients, see Table S10 in the Supplemental Material).

Association between corruption records and face-based trait inferences: further exploration of potential mechanisms. Finally, to elucidate whether the observed associations between trait judgments from faces and records of violations of the facial identities might in part be attributable to unintended properties of photo sources, we performed GLMM analyses on two subsets of data, respectively. For one subset of data, we excluded officials whose photos were not self-selected—that is, we included only officials whose photos were from government websites and personal campaign websites ($N = 69$; 33 violated campaign finance laws). Trait inferences based on photos self-selected by the officials were significantly associated with the officials' records of violations—corruptibility inferences: $b = 0.23$, $SE = 0.03$, 95% CI = [0.17, 0.29], $z = 7.48$, $p < .001$; dishonesty inferences: $b = 0.26$, $SE = 0.03$, 95% CI = [0.20, 0.32], $z = 8.68$, $p < .001$; selfishness inferences: $b = 0.25$, $SE = 0.03$, 95% CI = [0.19, 0.31], $z = 8.37$, $p < .001$; trustworthiness inferences: $b = -0.25$,

SE = 0.03, 95% CI = [–0.31, –0.19], z = –8.31, p < .001; and generosity inferences: b = –0.25, SE = 0.03, 95% CI = [–0.31, –0.19], z = –8.15, p < .001.

To test the hypothesis that potential negative biases in photos of officials with violations if the photos were taken after the violation was caught might be driving the face–corruption-record association, we performed GLMM analyses on a subset of data that included only officials for whom the dates on which their photo was taken was known (and were taken prior to the date when the violation was caught, for officials with violations; n = 75; 35 violated campaign finance laws). The associations between trait inferences and records remained significant: corruptibility inferences, b = 0.25, SE = 0.03, 95% CI = [0.19, 0.31], z = 8.33, p < .001; dishonesty inferences, b = 0.29, SE = 0.03, 95% CI = [0.24, 0.35], z = 9.78, p < .001; selfishness inferences, b = 0.29, SE = 0.03, 95% CI = [0.23, 0.34], z = 9.56, p < .001; trustworthiness inferences, b = –0.28, SE = 0.03, 95% CI = [–0.33, –0.22], z = –9.27, p < .001; and generosity inferences, b = –0.30, SE = 0.03, 95% CI = [–0.36, –0.24], z = –10.17, p < .001.

The analysis of the correlation between individual-level accuracies and response times was also planned. Results are not detailed here because these analyses intended to answer a question that is beyond the scope of the current article. For readers interested in these results, all relevant data and analysis codes can be accessed at https://osf.io/k4mds/.

Study 3

Study 2 replicated the face–record association found in Study 1 with an independent set of stimuli. However, these findings were based on traits that either were close in meaning to corruptibility (selfishness, dishonesty) or have the opposite meaning from it (trustworthiness, generosity). This resulted in our findings deriving from a single underlying factor with no comparison to different traits. Study 3 therefore aimed to test that the effects found in Studies 1 and 2 could be attributed specifically to corruptibility judgments.

Method

Participants. This study was preregistered before data collection began (https://osf.io/7a7eu/). To ensure a sample size comparable with that used in Study 1, we recruited 100 participants via MTurk. The same inclusion and exclusion criteria as in Study 1 were applied; in addition, participants were required to have no prior participation in Study 1. We excluded 15 individuals, 2 for not being native English speakers, 2 for pressing the same response key for all trials in a block, and 11 for failing to input valid responses for more than 10% of the trials in a block. After exclusions, data were retained from 85 participants who were recruited from MTurk in February and March 2017 (42 female; age: M = 37 years, SD = 10; 88% White, 6% Black, 4% Asian).

Stimuli and procedure. We used stimuli identical to those from Study 1 and a protocol similar to that of Study 1 except that participants evaluated the officials on a different set of traits: corruptibility, aggressiveness, masculinity, competence, and ambitiousness.

Results

Reliability of face-based trait inferences. We excluded from further analysis any responses faster than 100 ms and responses for officials who were recognized. Among the 85 participants, 3 recognized at least one official (in total, two officials were ever recognized).

We first checked the variation of individual-level ratings across faces for each trait. For all the four traits except masculinity, the majority of participants used the full scale to rate the faces, and over 92% of the participants used scores on both sides of the midpoint to rate the faces (see Fig. S7 in the Supplemental Material). Not surprisingly, given that all the officials were male, ratings for masculinity were skewed toward masculine; however, 80% of the participants still rated the faces on masculinity using scores on both sides of the midpoint.

Participants showed high consensus on face-based trait judgments for corruptibility, ICC = .86, F(52, 4368) = 8.7, 95% CI = [.80, .91]; aggressiveness, ICC = .85, F(54, 4536) = 8.3, 95% CI = [.79, .90]; masculinity, ICC = .89, F(53, 4452) = 13.6, 95% CI = [.85, .93]; and competence, ICC = .84, F(58, 4872) = 8.4, 95% CI = [.78, .89]; and the consensus on ambitiousness judgments was fair, ICC = .69, F(53, 4452) = 3.9, 95% CI = [.57, .79]. ICCs were computed according to type ICC(2, k) based on complete cases.

Association between corruption records and face-based trait inferences. Critically, we replicated the results found in Study 1 with this new set of participants: Officials who were convicted of political corruption looked more corruptible than their peers with clean records, aggregate-level accuracy = 72.22%, lower 95% CI = 62.12%, $\chi^2(1)$ = 13.35, p < .001; average individual-level accuracy = 56.30%, SD = 7.22%, lower 95% CI = 55.00%, $t(84)$ = 8.04, d = 0.87, p < .001. Additionally, participants in Study 1 and the present study viewed the same set of stimuli, and their judgments (averaged over participants within each study) of how corruptible a face looked were highly correlated, ρ = 0.88, 95% CI = [0.81, 0.92], p < .001.

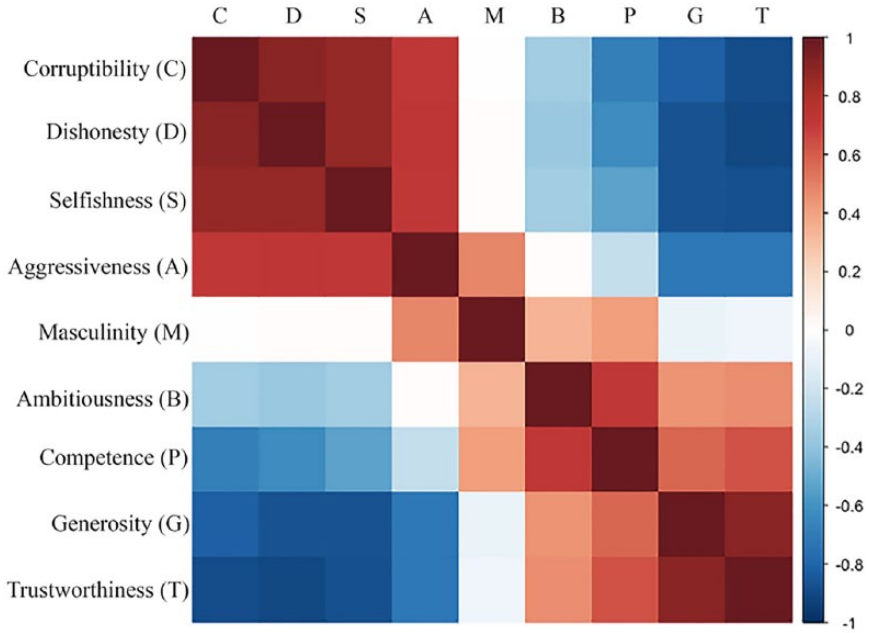

Fig. 2. Spearman correlation coefficients between each pair of traits across Study 1 and Study 3, calculated with aggregate-level trait ratings ($N = 72$). Inferences of corruptibility were averaged over the two studies.

Interestingly, data from the present study revealed that officials who were perceived as more aggressive were also more likely to have been convicted of political corruption, aggregate-level accuracy = 66.67%, lower 95% CI = 56.36%, $\chi^2(1) = 7.35$, $p = .003$; average individual-level accuracy = 55.09%, $SD = 6.13\%$, lower 95% CI = 53.98%, $t(84) = 7.66$, $d = 0.83$, $p < .001$. However, the associations between corruption records and inferences of masculinity, ambitiousness, and competence were not statistically reliable (95% CIs included 50%, and *p*s were > .01 for aggregate-level accuracies).

Correlation structure of trait inferences. Our primary interest in the current study was whether the observed face–corruption-record associations resulted from inferences of specific traits or global valence evaluations of the face. We first analyzed the correlation structure of the trait inferences. To allow for analyses across all nine traits (those from Study 1 and Study 3 combined), we first averaged inferences of traits across participants for each face, and then these aggregate-level data were merged across Study 1 and the present study. Figure 2 shows the Spearman correlation coefficients between each pair of traits. All correlations were in expected directions and generally strong, except for masculinity and ambitiousness.

A principal component analysis with varimax rotation indicated that these trait inferences clustered on three distinctive factors: a corruptibility-related factor (corruptibility, dishonesty, selfishness, aggressiveness, generosity, and trustworthiness), a competence-related factor (competence and ambitiousness), and a masculinity-related factor (masculinity), each accounting for 57%, 19%, and 15% of the variance in the data, respectively (see Table S11 in the Supplemental Material). A composite score was computed for each factor with the trait inferences that comprised it (Todorov et al., 2005; for the corruptibility-related factor, positive and negative traits were aggregated with opposite signs). Importantly, logistic regression analyses with each of these three factors independently while controlling for other covariates demonstrated that only the corruptibility-related factor was associated with corruption records (Fig. 3).

Study 4

Study 3 demonstrated that elected officials' corruption records were associated with specific trait inferences (e.g., corruptibility). A final important question concerns the facial features that make some officials look more corruptible than others. Study 4 provided a preliminary exploration of this question by first estimating the relationship between objective facial structures, inferences of traits, and officials' records with causal mediation models (Study 4a; not preregistered). Second, the causal effects suggested by the mediation analysis were directly

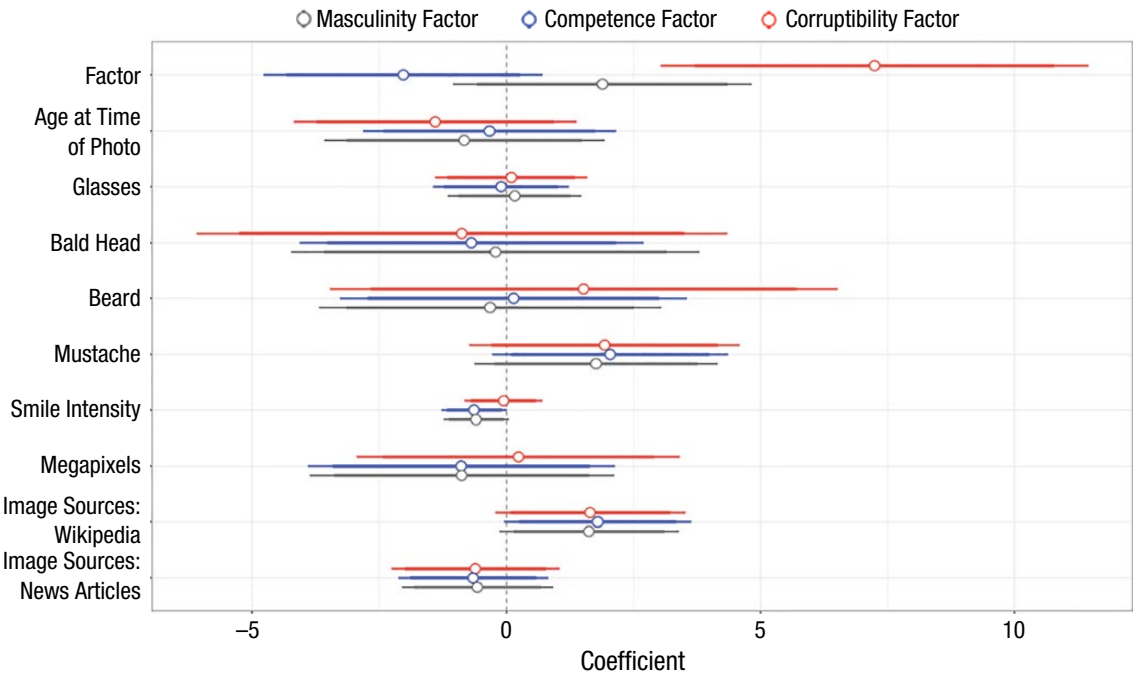

Fig. 3. Unstandardized logistic regression ($N = 72$) coefficients for factors and photo characteristics as regressors of the officials' corruption records (1 = conviction, 0 = clean) in Study 3. Thick lines represent 90% confidence intervals, and thin lines represent 95% confidence intervals. Glasses is a dummy variable with 1 indicating the official wore glasses. Bald head is a dummy variable with 1 indicating the official was bald. Beard is a dummy variable with 1 indicating the official had a beard. Mustache is a dummy variable with 1 indicating the official had a mustache. Smile intensity was coded manually with three levels (1 = smile with no teeth exposed, 2 = smile with teeth but not gums exposed, 3 = smile with gums exposed). There were three sources of photos: government and campaign websites (benchmark), Wikipedia, and news articles. All variables were normalized into the range of [0, 1].

tested in an experiment that manipulated the face stimuli (Study 4b; preregistered).

Study 4a

Method. Study 1 and Study 2 collected judgments of a common set of traits (corruptibility, dishonesty, selfishness, trustworthiness, and generosity) for two distinct sets of officials. The present study merged data from both studies. For trait judgments of an official given by a participant, we computed a composite score using his ratings across the five traits and referred to it as *corruptibility-related trait inferences*.

Officials were those used in Study 1 and Study 2. Whether an official is corrupt was measured by his record. A record of conviction of political corruption or violation of campaign finance laws suggests that an official is corrupt, and a clean record suggests an official is not corrupt. Officials' records are one metric of real-world corruption, but the potential measurement error of this metric is beyond the scope of the present study.

Eight metrics representing the distances between facial landmarks specified by anthropometric definitions were measured (Farkas, 1994; Stirrat & Perrett, 2010). Stimuli were the photos of the elected officials used in Study 1 and Study 2. We adjusted for shifts in posture or tilts in head angle by making all measurements only on one side of the face—the side turned most toward the camera—and by generating a face-based reference frame—the horizontal axis of the face was defined by the line connecting the two pupils, and the vertical axis was defined by the line through landmark *n* (for nasion) that was perpendicular to the horizontal axis (Fig. 4). Summary statistics of these metrics are reported in the Supplemental Material (Table S12).

Results. Studies 1 to 3 demonstrated that officials who had clean records were judged differently on corruptibility than officials who were convicted of political corruption and those who violated campaign finance laws. To test the hypothesis that the perceptual difference was mediated by certain facial structures, we analyzed a causal mediation model linking whether an official is corrupt, corruptibility-related trait inferences, and each facial structure with data from Study 1 and Study 2 (Fig. 5). The effect of whether an official is corrupt on the facial structure (path *a*) and the effect of the facial structure on

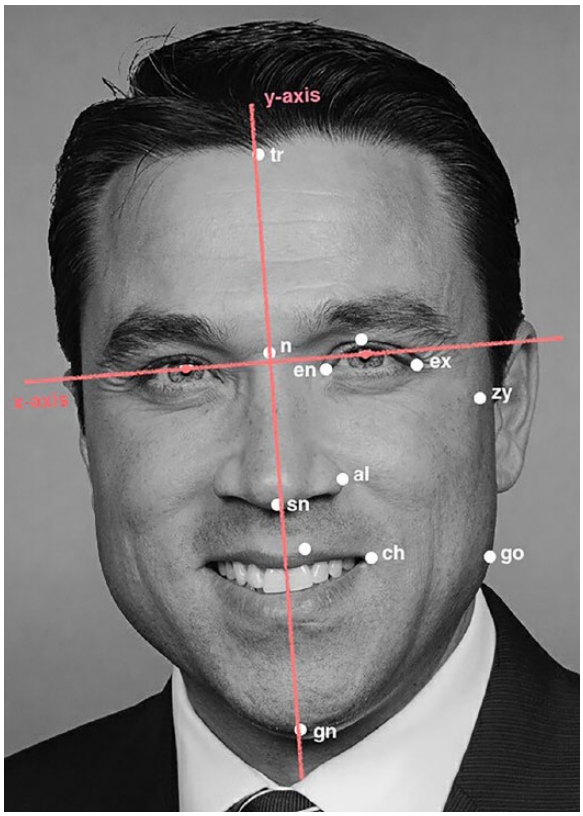

Fig. 4. Illustration of facial landmarks (white points) and the coordinate system (red lines). Facial width-to-height ratio was calculated as the bizygomatic width (the horizontal distance from landmark *zy* to the *y*-axis multiplied by 2) divided by the upper-face height (the vertical distance from the highest point of the upper lip to the highest point of the eyelids). Face width/lower-face height was calculated as the bizygomatic width divided by the lower-face height (the vertical distance between landmark *ex* and landmark *gn*). Lower face/face height was calculated as the lower-face height divided by the physiognomic face height (the vertical distance between landmark *tr* and landmark *gn*). Cheekbone prominence was calculated as the bizygomatic width divided by the jawbone width (the horizontal distance from landmark *go* to the *y*-axis multiplied by 2). Internal eye-corner distance was calculated as the ratio of the internal eye-corner width (the horizontal distance from landmark *en* to the *y*-axis multiplied by 2) to the bizygomatic width. Nose height was calculated as the ratio of the nose length (the vertical distance from landmark *n* to landmark *sn*) to the lower face height. Mouth width was calculated as the ratio of the mouth corner distance (the horizontal distance from landmark *ch* to the *y*-axis multiplied by 2) to the jawbone width. Nose/mouth width was calculated as the ratio of the nose width (the horizontal distance from landmark *al* to the *y*-axis multiplied by 2) to the mouth corner distance.

corruptibility-related trait inferences controlling for whether an official is corrupt (path *b*) constitute the indirect effect from whether an official is corrupt to corruptibility-related trait inferences (path *ab*). Path *a* was estimated with linear regression models. Path *b* was estimated with linear mixed models in which subjects, images nested within record types, and the interactions between subjects and record types were treated as random factors. The indirect effect was estimated with *RMediation* in R. The direct effect (path *c*′) of whether an official is corrupt on corruptibility-related trait inferences after controlling for the indirect effect was estimated in the same model as for path *b*. Photo characteristics (the official's age and smile intensity; the presence of glasses, a beard, a mustache, and a bald head; image clarity; and image sources) were included as covariates in all models; for simplicity, these paths are not depicted in the figure.

Two of the eight facial structures were identified to have significant indirect effects: facial width-to-height ratio (unstandardized coefficient for path *ab* = 0.06, *SE* = 0.03, 95% CI = [0.01, 0.12]), and face width/lower face height (unstandardized coefficient for path *ab* = 0.11, *SE* = 0.04, 95% CI = [0.04, 0.18]). These results revealed that compared with officials who had clean records, those who were convicted of political corruption and violated campaign finance laws were perceived more negatively (more corruptible, dishonest, and selfish and less trustworthy and generous), and these negative impressions were partially attributable to higher facial width-to-height ratio and face width/lower-face height.

Study 4b

Study 4a suggests that compared with officials with slimmer faces, officials with wider faces were judged more negatively on corruptibility-related traits. This finding raises an important question: Given the same elected official, is how corruptible he looks influenced by how wide his face is in a photo? Study 4b directly tested this causal hypothesis by manipulating the facial width of the photos and contrasting the degree of corruptibility inferred from the slim, original, and fat version photos of the same official.

Method.

Stimuli. This study was preregistered before data collection began (https://osf.io/58x6e/). Stimuli were 450 black-and-white headshots of real elected officials. There were three versions of the stimuli: original, fat, and slim. Original stimuli consisted of 71 photos from Study 1 and 79 photos from Study 2 (1 photo from Study 1 and 1 photo from Study 2 were excluded from the present study because the manipulation of face width distorted these two faces). These 150 original stimuli were further manipulated with the Adobe Photoshop Face-Aware Liquify tool to increase face width by 7% and decrease face width by 7%, which resulted in two additional versions of each facial identity (see Fig. 6 for an example; all stimuli used in the present study can be accessed at https://osf.io/k4mds/). Fat stimuli consisted of the 150 photos with increased face width, and slim stimuli consisted of the 150 photos with decreased face width.

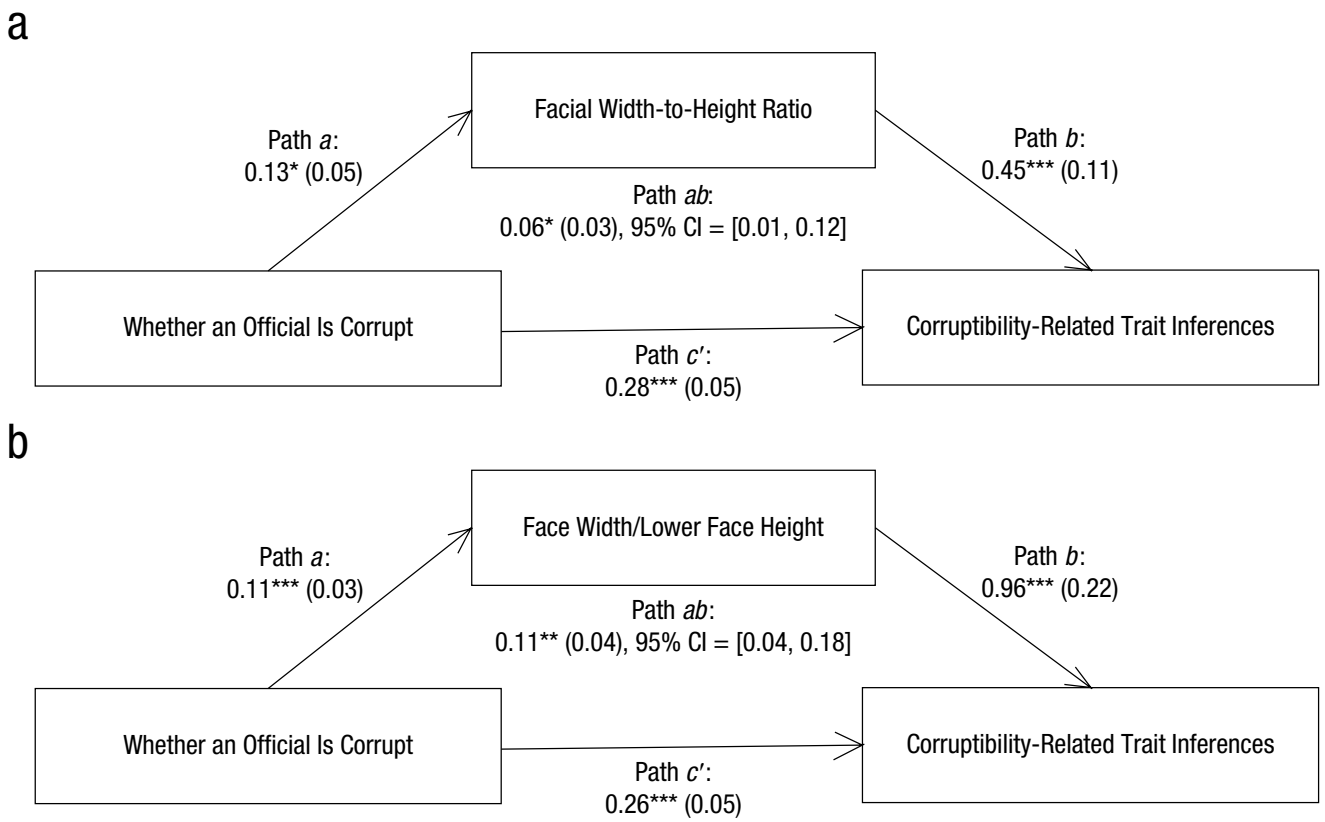

Fig. 5. Results of causal mediation analyses showing the influence of whether an official is corrupt on corruptibility-related trait inferences, as mediated by facial structures (Study 4a). A mediation model was constructed for each of the eight facial metrics separately and was tested with data from Study 1 and Study 2. Two of the eight facial metrics, (a) face width-to-height ratio and (b) face width/lower-face height, showed significant indirect effects. Unstandardized coefficients are shown, and standard errors are given in parentheses. Coefficients for path *a* were estimated in linear regression models. Coefficients for path *b* and path *c′* were estimated in linear mixed models. The indirect effects of path *ab* were estimated with *RMediation* in R. Photo characteristics were included as covariates in all models; for simplicity, these variables and the corresponding paths are not depicted in the figure. No indirect effect was found for the other six facial metrics. Asterisks indicate significant paths (*$p < .05$, **$p < .005$, ***$p < .0005$). CI = confidence interval.

This percentage of face-width change was the maximum manipulation we could achieve subject to the constraints that all faces should look natural and the manipulation should be subtle enough to go unnoticed.

Participants. To investigate our main hypothesis that the same official would be judged as more corruptible when his face was fatter relative to when it was slimmer, we conducted a pilot study on MTurk with 18 participants, which yielded 2,700 observations. These observations gave an estimated effect size of 0.09, justifying a minimum sample size of 16 participants and 2,375 observations. To ensure sufficient power even with data exclusion, we predetermined the sample size to be 100 participants. Participants were required to be located in the United States, to be 18 years old or older, and to have normal or corrected-to-normal vision, an educational attainment of high school or above, a HIT approval rate greater than or equal to 95%, and no prior participation in the pilot study. Additionally, two open-ended questions and one closed-ended question in the survey at the end of the experiment (Table S13 in the Supplemental Material) gauged whether participants noticed that the width of the faces was manipulated; participants who mentioned face width to any of the open-ended questions were excluded from data analysis.

Only 1 participant recognized that the face width of the stimuli was manipulated; this individual was excluded from data analyses. Another 19 participants were excluded for failing to input valid responses for more than 10 trials. After exclusions, the final sample consisted of 80 participants (37 female; age: $M = 38$ years, $SD = 10$; 89% White, 5% Black, 5% Asian), who were recruited from MTurk in July 2017.

Procedures. Participants were not informed about the purpose of the study (they were told only that this was a study about judging how corruptible politicians looked

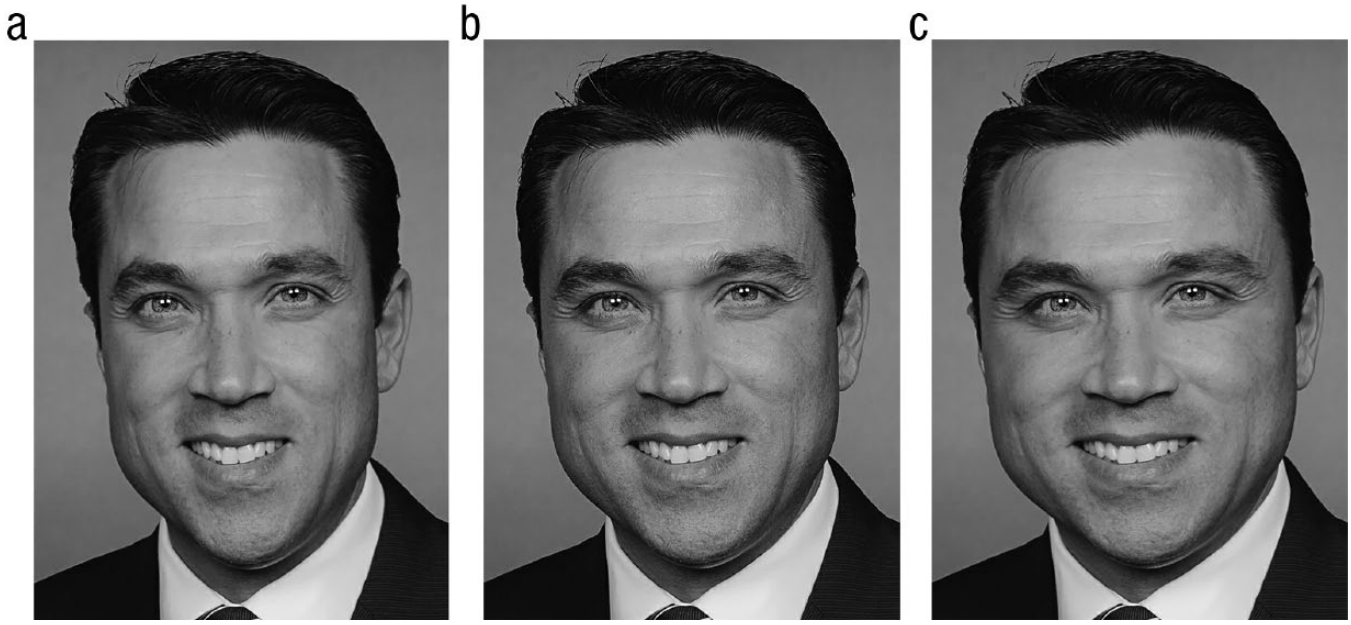

Fig. 6. Example of the same face in (a) slim, (b) original, and (c) fat versions.

on the basis of their photos). They were instructed to make their decisions as quickly and precisely as possible. Participants viewed and evaluated the 450 stimuli one at a time in 10 blocks; they had the option to take breaks between blocks. The order of the 450 stimuli was randomized under a constraint that different versions of photos of the same facial identity did not appear within 10 consecutive images. Participants indicated how corruptible each face looked on a 9-point Likert scale anchored with bipolar adjectives (*corruptible* and *incorruptible*) and were encouraged to use the full range to rate the faces. The orientation of the scale was randomized for each participant. Except for the change of the Likert scale, the present study followed the same experimental procedures as the previous studies. After rating all stimuli, participants were asked whether they had recognized any of the officials or noticed that the width of the faces were manipulated (see Table S13) and filled out a short survey questionnaire on demographic characteristics, political attitudes, and personality.

Results. The survey responses revealed that when told that photos of the same politicians were shown more than once during the experiment and asked whether photos of the same politician were used, 61 of the 80 participants reported that they believed the repeated photos of the same politician were identical. The rest of the participants either indicated that they were not sure whether the photos of the same politician were different or identical ($n = 3$) or mentioned that the faces in these photos might have different facial expressions ($n = 4$), hair or facial hair ($n = 4$), smile intensity ($n = 3$), eyes ($n = 1$), glasses ($n = 1$), or head shapes ($n = 1$); the individual pictured might be wearing different clothing ($n = 1$); or some photos looked scarier ($n = 1$), might be mixed with parts from other pictures ($n = 1$), or might be taken from different angles ($n = 1$).

Most of the participants (71.25%) used the full range to rate the faces as instructed, and all participants used both sides of the scale to rate the faces. Data were analyzed in linear mixed models, and subjects, images, and the interactions between subjects and versions were treated as random factors. As hypothesized, individual-level data showed that face width had a significant effect on inferences of corruptibility; specifically, a participant perceived an official as more corruptible when his face was fat relative to when his face was slim, $b = 0.06$, $SE = 0.02$, 95% CI = [0.03, 0.09], $p < .001$, $d = 0.22$. This unconscious perceptual bias was symmetrically driven by increasing face width, $b = 0.06$, $SE = 0.02$, 95% CI = [0.02, 0.11], $p = .008$, $d = 0.22$, and decreasing face width, $b = 0.06$, $SE = 0.03$, 95% CI = [0.02, 0.11], $p = .025$, $d = 0.22$. We further analyzed whether the perceptual bias to rate the fat version of a face as more corruptible than the slim version of that face varied as a function of the baseline corruptibility rating of the original photo, as planned in our preregistration. The ratings for each official in each version of the photo were first averaged over participants, and then these aggregate-level ratings for different photos of the same official were used to calculate perceptual biases. We did not observe significant correlation

between perceptual bias (fat vs. slim versions of the photo) and the corruptibility inferences based on the original version of the photo, $\rho = 0.01$, $SE = 0.03$, 95% CI = [–0.06, 0.07], $p = 0.789$ (photo characteristics were included as control variables; see Fig. S8 in the Supplemental Material).

General Discussion

Across three preregistered studies, we found evidence supporting the hypothesis that trait-specific inferences, such as corruptibility, made from photographs of officials' faces are associated with real-world measures of political corruption and violation. This association was replicated across officials at different levels of government. It was not driven by just a small subset of faces or fully explained by other photo characteristics, such as smile intensity. The association remained robust when analyses controlled for heterogeneous beliefs about corruption base rates and potential photo-selection biases.

It is important to distinguish accuracy as defined by agreement with consensus judgments from accuracy related to actual real-world metrics (Funder, 1987). Similar to prominent studies of the association between competence judgments and election success (e.g., Todorov et al., 2005), our present work has pursued the latter interpretation of accuracy. The accuracy related to corruption records we found was comparable with that related to election success—for instance, Todorov et al. (2005) found that for 2004 U.S. Senate races, aggregate-level accuracy was 68.8%, and average individual-level accuracy was 53%. We emphasize that for our present work and a large literature on the association between face-based trait judgments and real-world metrics, accuracies at an individual level were only slightly above chance (but significantly so), and participants were very often wrong. However, the considerably larger effect sizes for aggregated judgments have important implications for real-world collective decisions such as elections and corruption investigations.

In Study 4, we found that an official was perceived as more corruptible when his face was manipulated to be slightly wider and less corruptible when his face was manipulated to be slightly slimmer, even though participants did not detect such manipulation of the facial identity. Our finding dovetails with the large literature on perceptual biases related to face width-to-height ratio (e.g., Deska et al., 2018) and the literature on weight stereotypes, which shows that overweight individuals are judged as lazy, greedy, selfish, and less trustworthy (Greenleaf, Chambliss, Rhea, Martin, & Morrow, 2006; Larkin & Pines, 1979). Yet widening or narrowing the face potentially introduces other changes to the geometry of the face. It will be important for future studies to investigate which of these correlated structural changes are in fact detected by the brain and drive the change in social judgments that perceivers make.

The detailed causal mechanisms that ultimately underlie the association between a record of corruption and face-based judgments of corruptibility we found are likely to be complex and bidirectional (Swann, 1984). In particular, people who look corruptible might be more likely to be approached by others with the intent to corrupt them, which in turn results in the mutual behaviors required for corruption to occur (Kruglanski, 1989); further experimental studies would be required to tease apart their relative contributions.

Given these considerations, we emphasize that our findings should be interpreted with caution. Do they show that corruptible individuals have a different facial structure, as suggested by physiognomy? There are strong reasons to be skeptical. First, the record of an official is unlikely to be an errorless measure of how corruptible he actually is. Second, the photographs posted on government and campaign websites might provide a biased representation of an official's face—for example, some photos have clearly been retouched for skin texture and lighting. Third, there might well be other unknown confounding effects. For example, perhaps officials who committed a corrupt act might have looked more guilty when they posed for photos (even prior to being convicted), which in turn could have provided subtle visual cues for trait judgments. Fourth, both face judgments and social behaviors strongly depend on context (Todorov, 2017). For instance, in the context of business corruption, business executives' corruption records were not associated with how trustworthy they looked (Rule, Krendl, Ivcevic, & Ambady, 2013). These findings and the considerations mentioned previously suggest that the self-fulfilling prophecy (e.g., Haselhuhn et al., 2013; Slepian & Ames, 2016) together with biases in judicial decisions (e.g., Wilson & Rule, 2015; Zebrowitz & McDonald, 1991) may be more plausible explanations than physiognomy for the face–corruption-record association we found. Future studies should examine multiple photographs of the same official taken in different contexts (e.g., posed and candid photos) and at different time points (e.g., from the time he ran for office, after misconducts, and after prosecutions).

There are important limitations to the generalizability of our studies (Simons, Shoda, & Lindsay, 2017). They leave open whether such trait judgments operate in the real world, where faces are not photographs and whether similar associations would hold for other cultures or for antisocial behaviors among people in general. It is also possible that corruptibility judgments are better

correlated with other social behaviors that we did not measure, which in turn provide an indirect link to recorded corruption—indeed, it is conceivable that prosecutors' decisions might be one such social behavior.

We conclude with a future direction suggested by this work. The ultimate explanation for the findings we report must reside in evolutionarily based or experience-based neural mechanisms in the brains of both the subjects making the social judgments and the officials engaging in the corrupt behaviors. For instance, there are already claims that individual differences in traits can be predicted from patterns of brain activity (Dubois, Galdi, Paul, & Adolphs, 2018; Finn et al., 2015), and there is a large literature showing that social inferences engage specific neural networks in the brains of perceivers (Spunt & Adolphs, 2017). Future studies using neuroimaging could help further uncover the causal mechanisms behind our findings.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

All authors developed the study concept and designed the study. Testing and data collection were performed by C. Lin, who also analyzed and interpreted the data under the supervision of R. Adolphs and R. M. Alvarez. All authors drafted the manuscript. All the authors approved the final manuscript for submission.

Acknowledgments

We thank Colin F. Camerer, Antonio Rangel, Anita Tusche, and Shuo Wang for helpful conversations.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported in part by a Conte Center grant from the National Institute of Mental Health (P50MH094258).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618788882

Open Practices

All data and materials have been made publicly available via the Open Science Framework (OSF) and can be accessed at https://osf.io/chpfn/. The design and analysis plans for the experiments were preregistered at the OSF (Study 1: https://osf.io/mge8r/, Study 2: https://osf.io/tgzpz/, Study 3: https://osf.io/7a7eu/, Study 4: https://osf.io/58x6e/). The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618788882. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

References

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science*, *26*, 270–275.

Berry, D. S., & Zebrowitz-McArthur, L. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin*, *14*, 23–33.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*, 674–679.

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, *25*, 1132–1139.

Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018). Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *Journal of Personality and Social Psychology*, *114*, 75–94.

Dubois, J. C., Galdi, P., Paul, L. K., & Adolphs, R. (2018). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *bioRxiv*, Article 257865. doi:10.1101/257865

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*, 1508–1519.

Farkas, L. G. (Ed.). (1994). *Anthropometry of the head and face*. New York, NY: Raven Press.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., . . . Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*, 1664–1671.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.

Genevsky, A., & Knutson, B. (2015). Neural affective mechanisms predict market-level microlending. *Psychological Science*, *26*, 1411–1422.

Geniole, S. N., Denson, T. F., Dixson, B. J., Carré, J. M., & McCormick, C. M. (2015). Evidence from meta-analyses of the facial width-to-height ratio as an evolved cue of threat. *PLOS ONE*, *10*(7), Article e0132726. doi:10.1371/journal.pone.0132726

Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2017). Facial appearance affects science communication. *Proceedings of the National Academy of Sciences, USA*, *114*, 5970–5975.

Greenleaf, C., Chambliss, H., Rhea, D. J., Martin, S. B., & Morrow, J. R. (2006). Weight stereotypes and behavioral intentions toward thin and fat peers among White and Hispanic adolescents. *Journal of Adolescent Health*, *39*, 546–552.

Hamermesh, D. S. (2011). *Beauty pays: Why attractive people are more successful*. Princeton, NJ: Princeton University Press.

Haselhuhn, M. P., Wong, E. M., & Ormiston, M. E. (2013). Self-fulfilling prophecies as a link between men's facial width-to-height ratio and behavior. *PLOS ONE*, *8*(8), Article e72259. doi:10.1371/journal.pone.0072259

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review*, *93*, 429–445.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*, 395–409.

Larkin, J. C., & Pines, H. A. (1979). No fat persons need apply: Experimental studies of the overweight stereotype and hiring preference. *Sociology of Work and Occupations*, *6*, 312–327.

Lin, C., Adolphs, R., & Alvarez, R. M. (2017). Cultural effects on the association between election outcomes and face-based trait inferences. *PLOS ONE*, *12*(7), Article e0180837. doi:10.1371/journal.pone.0180837

Olivola, C. Y., Eastwick, P. W., Finkel, E. J., Hortacsu, A., Ariely, D., & Todorov, A. (2014). *A picture is worth a thousand inferences: First impressions and mate selection in Internet matchmaking and speed-dating*. Pittsburgh, PA: Tepper School of Business, Carnegie Mellon University.

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*, 83–110.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition*, *24*, 607–640.

Ravina, E. (2012). Love & loans: The effect of beauty and personal characteristics in credit markets. *SSRN*. doi:10.2139/ssrn.1107307

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLOS ONE*, *7*(3), Article e34293. doi:10.1371/journal.pone.0034293

Rose-Ackerman, S. (2013). *Corruption: A study in political economy*. San Diego, CA: Academic Press.

Rule, N. O., Ambady, N., Adams, R. B., Jr., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, *98*, 1–15.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*, 409–426.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Slepian, M. L., & Ames, D. R. (2016). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological Science*, *27*, 282–288.

Spunt, R. P., & Adolphs, R. (2017). The neuroscience of understanding the emotions of others. *Neuroscience Letters*. Advance online publication. doi:10.1016/j.neulet.2017.06.018

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science*, *21*, 349–354.

Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*, 457–477.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton, NJ: Princeton University Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623–1626.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545.

Valentine, K. A., Li, N. P., Penke, L., & Perrett, D. I. (2014). Judging a man by the width of his face: The role of facial ratios and dominance in mate choice at speed-dating events. *Psychological Science*, *25*, 806–811.

Valla, J. M., Ceci, S. J., & Williams, W. M. (2011). The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology*, *5*(1), 66–91.

Van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*, 796–803.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*, 1325–1331.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*, 603–623.